Statistical Techniques for Determining the Repeatability of Man-in-the-Loop System Performance Data

Captain Sandra K. Smith
JADS Joint Test Force
2050A 2nd St. S.E.
Kirtland Air Force Base, New Mexico 87117-5522
505-846-0462
smiths@jads.kirtland.af.mil

Captain Roman Nation
JADS Joint Test Force
2050A 2nd St. S.E.
Kirtland Air Force Base, New Mexico 87117-5522
505-846-0939
nation@jads.kirtland.af.mil

Major Darrell L. Wright
JADS Joint Test Force
2050A 2nd St. S.E.
Kirtland Air Force Base, New Mexico 87117-5522
505-846-1015
wright@jads.kirtland.af.mil

ABSTRACT: *The Joint Advanced Distributed Simulation (JADS) Joint Test Force (JTF) was chartered by the Office of the Secretary of Defense to investigate the utility of advanced distributed simulation (ADS) technology to test and evaluation (T&E). JADS executed three test programs (command, control, communications, computers and intelligence [C4I]; precision guided munitions; and electronic warfare) representing slices of the overall T&E spectrum, as well as observing other activity within the T&E community, to form its conclusions. One of the slices, the Electronic Warfare (EW) Test, compared system performance data collected using ADS to data collected in a stand-alone hardware-in-the-loop (HITL) facility and on an open air range (OAR) via traditional test methods. In order to isolate and evaluate the impacts of distribution from other potential variance sources, testing was conducted under tightly controlled conditions. Yet, because of the statistical nature of the performance of man-in-the-loop systems, noise in electronic systems, and difficulty in replicating human and environmental conditions, it was extremely difficult to repeat a particular test without obtaining slightly different results. This paper discusses the application of statistical analysis techniques to evaluate the repeatability, or consistency, of the EW Test data collected within individual test facilities, in order to aid in the identification and understanding of the impacts of the variance sources within, and across, test phases.*

1.0 INTRODUCTION

In testing the performance of man-in-the-loop systems, even under the most tightly controlled conditions, it is likely that the results of a particular test, when repeated, will differ slightly. This variation is due to the statistical nature of the performance of such systems, noise in electronic systems, and difficulty in replicating human and environmental conditions. These differences may be small in a well-established test facility, but nonetheless, they are non-zero.

For the recently completed Joint Advanced Distributed Simulation (JADS) Electronic Warfare (EW) Test, system under test (SUT) data was collected across four test phases for ten individual measures of performance (MOPs). Each test phase involved different equipment, operators, and facilities, in order to provide a capability for comparing data collected using Advanced Distributed Simulation (ADS) to data collected in a stand-alone hardware-in-the-loop (HITL) facility, in a system integration lab (SIL), and on an open air range (OAR) via traditional test methods. The evaluation process for each MOP included correlating equivalent data sets across test phases to identify areas where distributed test control, network performance, data loss, latency or other ADS-induced factors impacted the quality or validity of SUT performance data.

The correlation methodology involved using statistical hypothesis testing on distribution shape, location (mean) and dispersion (variance) parameters. For each parameter, an appropriate statistical comparison test was selected based on the distribution form of the collected data (e.g., binomial, normal.) The underlying hypothesis of each test was that the two data samples were equivalent; that is, that they represented the same true population. If, in performing the test, this hypothesis could not be rejected with reasonable confidence, then the two data sets were determined to correlate.

Generally, poor correlation was obtained between different phase data, and the impacts of slight differences in threat system representation or operator practices at the different facilities greatly overshadowed the impacts of distributed testing methods. Operator variance across runs within facilities, and process, equipment, or environment changes, were factors which contributed to the inability to obtain correlation between data sets from different phases and facilities.

In order to support these conclusions, further analysis determined whether the data collected from each individual test facility, or during each particular test phase, was consistent, or repeatable, since the ability to correlate data across phases presumed some level of underlying data consistency. EW test analysts sought techniques to determine acceptable levels of data variability, or acceptable ranges within which data points must fall. This indicated if test results from any facility or phase were indicative of standard performance. The repeatability analysis performed was based strongly on engineering assessment, using basic statistical analysis tools and procedures. These fundamental techniques were surprisingly effective for highlighting data inconsistencies, including trends and extreme outlying values and for indicating system or process changes.

2.0 EW TEST OVERVIEW

The tasking to conduct an ADS-based EW test called for an airborne self-protection jammer (SPJ) as the system under test (SUT). The emphasis of the EW test was on the performance of the ADS components and their contribution or impact to testing rather than on the performance of the SPJ . Measures of performance (MOPs) for the SPJ were identified as measures that would most likely be affected by distributed testing.

The EW test was designed as a three-phase effort. Conducted partly on the OAR and partly at the Air Force Electronic Warfare Evaluation Simulator (AFEWES) HITL facility in Fort Worth, Texas, Phase 1 established a baseline of environment and jammer performance data against two command-guided surface-to-air missile (SAM) sites, one semi-active surface-to-air missile site, and one anti-aircraft artillery (AAA) site. The reference test condition and baseline data were used to develop the ADS test

environment for the two subsequent ADS test phases and provided the baseline data for comparison with the ADS test results.

Phase 2 tested a real-time digital system model (DSM) of the ALQ-131 receiver processor linked with terminal threats at the AFEWES facility and a scripted model of the terminal threat hand-off portion of an integrated air defense system (IADS). The reference test condition used in the Phase 1 flights was replicated as closely as possible in the synthetic ADS environment; the DSM was used, via the scripted flight profiles developed from the actual open air range (OAR) baseline flights against the AFEWES threats positioned in the synthetic environment as the threats were positioned on the range.

Phase 3 was conducted using the ALQ-131 jammer installed on an F-16 aircraft located in the Air Combat Environment Test and Evaluation Facility (ACETEF) located at Patuxent River Naval Air Station, Maryland. This facility was linked with AFEWES threats using the same reference test condition as the previous tests and controlled by the same scripted flight profiles.

For both ADS phases, the High Level Architecture (HLA) was used to link the DSM, and installed jammer, respectively, located at ACETEF, to the man in the loop terminal threats at the AFEWES facility, and to other models hosted in the JADS test control facility. Components of the JADS test are shown in Figure 1.
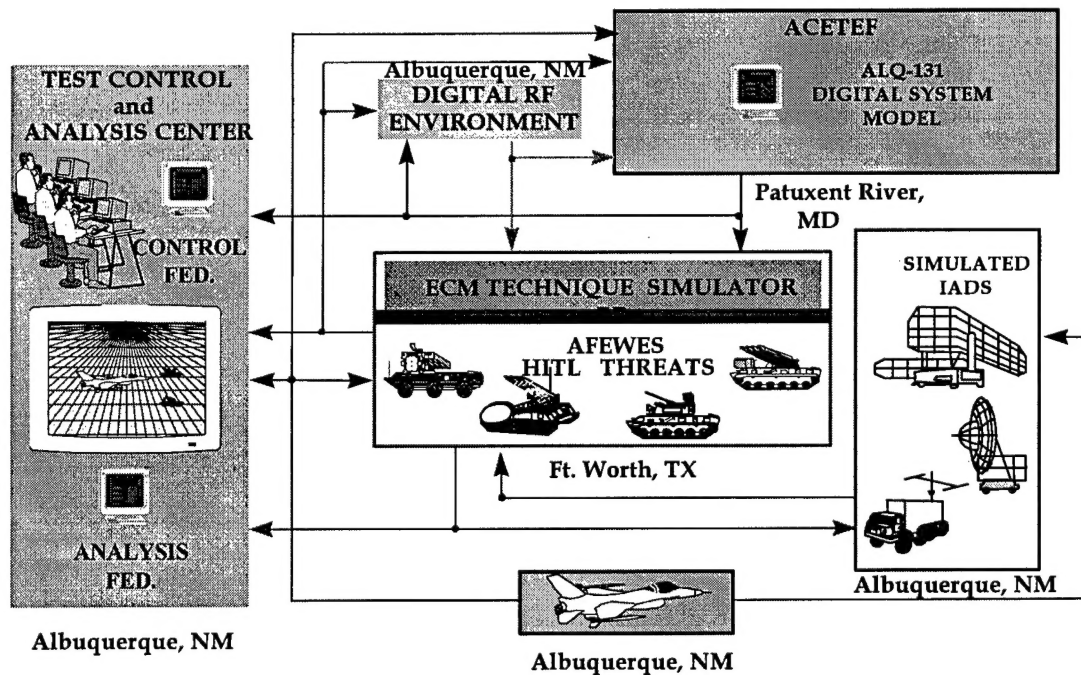
ECM = electronic countermeasures
IADS = Integrated Air Defense System
RF = radio frequency

Figure 1. JADS EW Test Components

The overall test approach was designed to provide a means of capturing the ADS effects within the Phase 2 and Phase 3 architectures. EW MOP data from each phase was collected and summarized; preliminary analysis included the generation of descriptive statistics and the determination of distribution shape and parameters via the use of frequency histograms. Finally, statistical comparison, using hypothesis testing, was used to correlate the EW MOP data sets, for each threat system and reference test condition, across

phases. The differences between data collected during ADS testing and baseline data collected in the OAR, SIL, and HITL phases were hoped to point JADS analysts to areas where ADS testing had significant impacts; likewise, similarities between ADS and non-ADS phase data were to confirm little ADS impact.

When unexpected variance sources within and across test facilities inhibited correlation, as well as the capability to attribute its lack to some ADS-induced factor, further analysis into data repeatability was performed. The remainder of this paper discusses the application of statistical analysis techniques to evaluate the repeatability, or consistency, of the EW Test data collected, to aid in the identification and understanding of the impacts of the variance sources within, and across, test phases.

3.0 MEASURE OF PERFORMANCE (MOP) DATA REPEATABILITY EVALUATION

For the EW test, repeatability analysis was accomplished after collected SUT performance data was correlated across phases to explain why some of the MOP data sets did not "match up" as expected. By providing insight into the variation that occurred between individual runs on a single day and for runs performed across test days at a single facility, repeatability analysis showed where data collected from an individual facility, or during a particular EW test phase, was not representative of "true" system performance. No specific rule determined what was "repeatable enough"; instead, the objective in analyzing each collected data set was to make an engineering assessment as to whether the sample faithfully characterized the true population of data it was collected to represent, or whether there was some process change or other source of variability that could be identified.

3.1 SUMMARY STATISTICS REVIEW

Given this goal, repeatability analysis of the EW SUT performance measure data was performed using a combination of evaluation techniques; including review of generated summary statistics, and visual assessment of plotted data versus time of collection. Each step was designed to highlight potential inconsistencies in the data. The first step was to identify any data points that did not fall within expected boundaries, as well as to identify sample data sets that did not follow an expected distribution (e.g., binomial, normal). This was accomplished through study of the data set's minimum and maximum values, as well as the range and variance of the data values collected. Analysts compared these sample statistics to expected boundaries to ensure that the sample range and distribution of values made logical sense for that performance measure. Unusual variance in the data, including extreme lack of variance or the existence of extreme outlying data points, was further researched to determine if anomalous system behavior occurred.

3.2 CONSISTENCY ASSESSMENT

The next step involved evaluating the consistency of the data collected over time. Changes in range or variance across runs or days pointed analysts to potential system, environment, or process changes, including operator learning. Visual analysis, through scatter plotting, or "run charting", of the data collected versus time, was an indispensable technique for examining the consistency of data behavior, and identifying any parameter values or trends that seemed "out of line" with the norm.

Trends and anomalies, such as outliers, or changes in average value or variability, were easily identified from such plots. Where inconsistencies or patterns in the data were found, analysts attempted to trace the source to some unusual test behavior by consulting facility subject matter experts and written documentation in test execution logs. Although analysts met with success in a several cases, it is important to note that not all unusual looking data could be explained.

3.3 TRUE POPULATION CHARACTERIZATION ASSESSMENT

As discussed above, there is no irrefutable guideline to follow in determining whether sample data is "good enough"; the data simply must be repeatable enough to allow confident statements to be made about the true population from which it comes. Since EW correlation analysis was performed using statistical hypothesis testing on distribution shape, location (mean) and dispersion (variance) parameters, any inconsistencies in the data that skewed these particular parameters usually negatively impacted correlation.

Thus, as a final step, EW analysts attempted to judge how well each true performance population had been characterized by the sample data collected. Two additional statistical parameters, the standard error of the mean and the sample size, were utilized.

The standard error of the mean is a statistic that characterizes the accuracy of the sample mean calculated for a data set, as an indicator of the true population mean. Based on the number of samples and variability, it is a measure of the distance on either side of the sample mean within which the true population mean should fall, with some particular likelihood. Sample size and variance are important aspects of this calculation, as the impact of inconsistent (highly variable) data behavior on a small set is greater than on a large one. Alternatively, more samples are required to gain confidence in parameters estimated from a high variance data sample than a low one. Standard error values can be computed for numerous calculated sample statistics, depending on the type and distribution of the data. The standard error of the mean is calculated as follows:

$$\text{standard error}_{mean} = \frac{s}{\sqrt{n}}$$

where: s = sample standard deviation (square root of variance); and
        n = number of samples

If the standard error is fairly small, then the true population mean likely falls very close to the calculated sample mean, and the "true" population of performance data has been fairly well characterized. Naturally, if the standard error is fairly large, the reverse is true, and conclusions drawn about the population from sample parameters may be inaccurate. Engineering assessment is required to determine what level of characterization is necessary, depending on the particular performance measure and how the data will be utilized. Anomalous data points, if judged to have occurred outside specified test conditions, can be excluded from the usable data set. This was done for a very small numbers of data point collected during the test phases.

4.0 RESULTS

Repeatability analysis for the EW Test was performed as described above, and a summary of the repeatability of the data collected for each MOP, for each threat system, and for each phase, was presented in EW Test report documentation. Presented here are some examples of data identified as repeatable, or non-repeatable, by EW Test analysts.

Figure 2 shows Threat System 2 ECM response time values collected over time during Phase 3, an excellent example of what repeatable data should look like when plotted versus time collected. Note that the occurrence of high and low values is consistent, and within expected range boundaries, across runs.
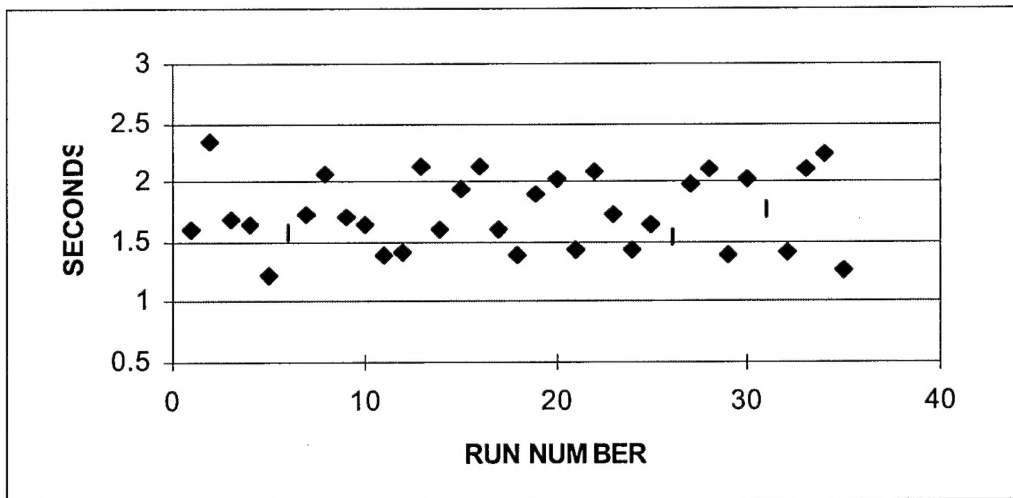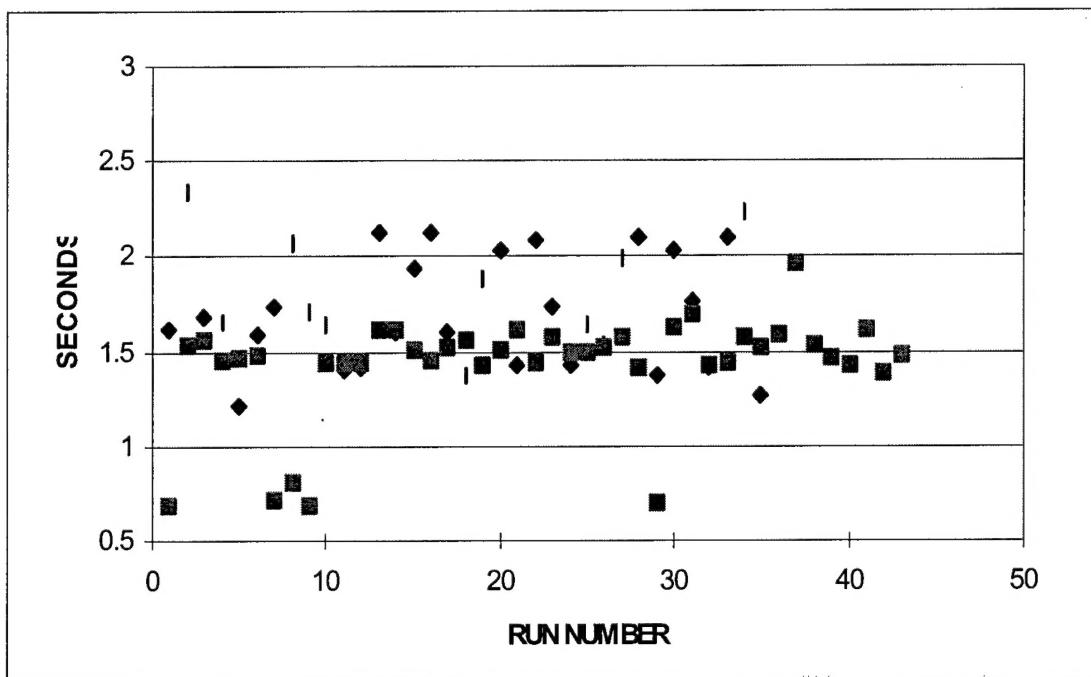
FIGURE 2: SYSTEM 2 ECM RESPONSE TIME - PHASE 3 SOUTH BOUND

Figure 3, however, shows data collected for the same threat during Phase 2 and indicates problems with repeatability. The range of collected data values is smaller than expected for this performance measure, and the outlying high and low values cast suspicion on characterized mean and variance values. Further research identified a process problem with the digital system model used in Phase 2 which resulted in the generation of an atypically tight stream of response time values. The cause of the outlying values was not determined.
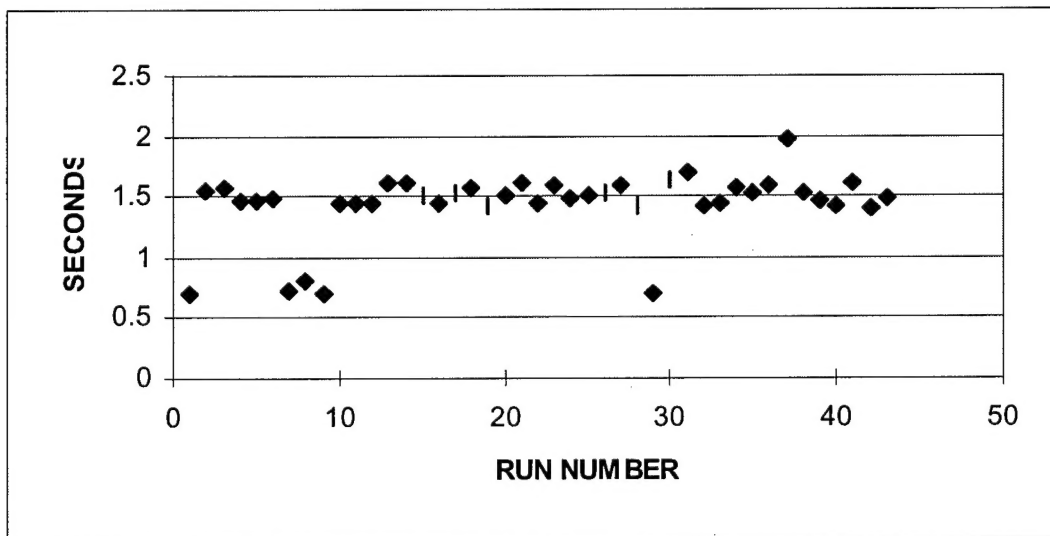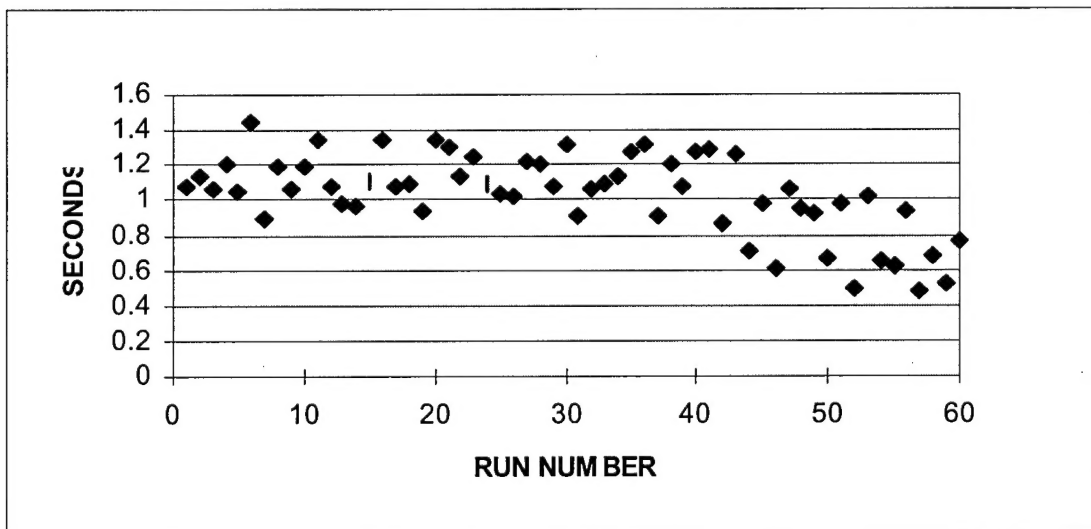

FIGURE 3: SYSTEM 2 ECM RESPONSE TIME - PHASE 2 SOUTH BOUND

Figure 4 shows the previous two data sets plotted together. It is obvious that the statistical parameters (i.e., mean, variance, and range) generated from each of these individual sets will not be similar enough to allow strong, if any, statistical correlation.

FIGURE 4: SYSTEM 2 ECM RESPONSE TIME - PHASE 2 SOUTH BOUND

Figure 5 shows ECM response time data collected over time for Threat System 1 at the SIL test during Phase 1. The lower range and average value of the last fifteen data points collected seems to highlight some change in test process or conditions. Further research identified a system change to repair an overheated component that was fixed later in the test day. Real-time assessment of response times during test execution did not identify the difference in the data samples being collected.

FIGURE 5: SYSTEM 1 ECM RESPONSE TIME - PHASE 1 SOUTH BOUND

5.0 CONCLUSIONS

In general, the EW Test data repeatability problems discovered were different for each threat system, phase, and MOP depending on the environment, the equipment involved, the ability of the operator to influence outcomes, and even the complexity of the measure. Yet, repeatability analysis, performed on

every data set collected using fairly basic statistical analysis tools and principles, gave EW Test analysts a notable amount of additional insight into the quality of collected SUT performance data.

As used post-test for EW, repeatability analysis techniques helped analysts identify data inconsistencies, track unknown variance sources, and further their understanding of the impacts of such sources on every phase of collected data. Although no EW Test data was thrown out based on the results of the repeatability analysis, some credible reasons for doing so were identified. Instead, heightened comprehension of variance factors was used in the interpretation of statistical correlation results, which gave explanation for the lack of correlation in many cases.

6.0 RECOMMENDATIONS FOR FUTURE TEST EFFORTS

Given their tremendous potential for providing insight into process changes, repeatability analysis techniques, as employed by the EW Test analysts, could be used as near real-time tools for determining the quality of SUT performance data as it is being collected, much as statistical process control techniques are used for quality assurance by the manufacturing community. Scatter plots, or run charts, could be used to ensure the consistency of collected data, highlighting system, environment, or process changes almost as soon as they occur. To implement this capability to future ADS tests, existing analysis tools must be adapted to plot run results over time, or new analysis tools must be developed to make real time repeatability analysis a reality.